Proceedings of the International Conference on Augmented Intelligence and Sustainable Systems (ICAISS-2022)
DVD Part Number: CFP22CB2-DVD; ISBN: 978-1-6654-8961-4

# Web Scraping Technique for Prediction of Air Quality through Comparative Analysis of Machine Learning and Deep Learning Algorithm

G.Kalaivani
Research Scholar,
Department of Computer Science,
School of Computing Sciences
Vels Institute of Science, Technology
& Advanced Studies (VISTAS)
Pallavaram, Chennai, India.
kalaivanikamalakannan@gmail.com

S. Kamalakkannan,
Associate Professor,
Department of Information Technology,
School of Computing Sciences
Vels Institute of Science, Technology
& Advanced Studies (VISTAS)
Pallavaram, Chennai, India.
kannan.scs@velsuniv.ac.in

**Abstract-- Air contamination has turned into a significant and difficult issue all over the planet and its direct impact with human well-being has drawn a lot of consideration from numerous analysts. Individuals are turning out to be known better ways of checking air quality data which are essential to safeguard human wellbeing from the genuine medical conditions brought about via air contamination. Numerous specialists are working on current air quality observation and expectations to carry out different government arrangements connected with the climate or air contamination and give precise outcomes to assist with settling on significant choices. This paper employs a machine learning method to implement predictive analytics and create a more accurate prediction model. These models are created by analysing trends and patterns using historical time series data and then creating a prediction model to forecast future values. These prediction models will be used to execute our suggested approach, the Air Quality Prediction Model (AQPM). This model yields a prediction model that accurately predicts the Air Quality Index (AQI) through the data collected. The information will be scraped from the Central Pollution Control Board (CPCB) website using the web scraping technique. The comparative analysis of ML and DL suggests that Long Short-Term Memory (LSTM) is the best fit model to measure air quality using three different accuracy metrics. Finally, the data are analysed using the predicted AQI in the LSTM model.**

***Keywords: Deep learning, LSTM, air quality index, prediction, CPCB, web scraping***

## I. INTRODUCTION

City residents are in increasing number and are aware of health and commercial costs of air contamination or pollution. Deprived air quality kills more than 4 million people every year. 97% of the public in big metropolises are visible to pollution echelons above endorsed parameters [1]. The rate of air pollution in cities is 3% of uncivilized native inventions in established countries and 6% in emerging countries [2]. For these motives, most metropolitan areas are increasingly focusing on real-time tracking of environmental and meteorological constraints [3]. In the past, the use of environmental statistics devices was primarily a government privilege. All gases in the atmosphere are damaging and even poisonous. Most of these are easy and can be prohibited cheaply. This research focuses on some of these gases and the dangerous hazards that are polluting the air. The most hazardous gases are carbon monoxide (CO), odourless, tasteless and colourless gases. There are several CO bases such as internal combustion engines, tandoors, and stoves. It also arises from the unfinished combustion of fuel. Carbon monoxide is one of the mutual poisonous gases in the world. It binds to Red Blood Cells (RBC) and is incapable to transmit oxygen to the body, which can prime to appropriations and death even at very low levels. The next hazardous gas is nitrogen dioxide (NO2). It is a communal air contaminant that produces millions of loads of loads each year. The pungent odour of gas is informal to evade, but it is still hazardous since even the short-term exposure can have long-term things. Acquaintance to this gas between thirty minutes and twenty-four hours can adversely affect breathing health. Airway swelling and augmented asthma signs in people suffering from it. Methane is odourless and extremely volatile at low concentrations and can cause choking. Another inflammable gas is hydrogen, which can be very fiery at certain absorptions. While these gases are a potential threat to all homes, shielding from these airs (through firms that provide protective devices and sensors) is expensive. The data are collected from the CPCB website and use the web scraping technique to scrap the data consisting of city, date, and seven required air pollutants. Through that AQI will be predicted and analyzed through Recurrent Neural Networks (RNN) models. Collecting data from a very partial number of radars throughout the city means that these standards are city-wide, date-wise, and the seven required air pollution. In fact, if we are getting air quality data on a busy street relatively more than in the yard, such as the posterior of a community on the same street, it makes sense that some values make a big difference. In fact, it is

affordable to have massive differences in a few values whilst the air first-rate facts are measured on a high visitor's road rather than on a lawn, as an example, within the return of a residence located at the same venue. Of late, there's a deeper understanding of the conservational parameters Particulate Matter (PM1.0), PM2.5, Carbon Monoxide (CO), Carbon dioxide (CO2), Sulphur dioxide (SO2), Ozone(O3), Hydrogen Sulphide (H2S), Nitrogen Dioxide (NO), and Nitrogen Monoxide and Nitrogen Dioxide (NOx) and how they are affected by the structure of the city, the causes of high levels of contaminants and misters, high-level registration and its distribution/diffusion dynamics etc. Therefore, the city appropriately regulates the mobility and other activities of the city, and the users of the city attach importance to the technology informed, as realized in [4], and the health and life of the citizens. We recognize that we live in a city that focuses on quality. The main objectives of the paper are as follows:

- To implement web scraping technology to collect real-time Air Quality Index (AQI) datasets from your website and manage them in a database.
- To determine data mining and data visualization techniques to extract information and insights from your dataset.
- To compare and analyse different ML and DL models to predict AQI.
- To implement and improve the accuracy of the model and use test data to measure and predict the accuracy of the best fit models.

The rest of the paper is organized as: Section II defines an overview of the relevant literature for existing systems. Section III describes the approaches and techniques used in the analytical part. The fourth section defines the proposed approach and complete results on the AQPM, and the fifth section describes the performance analysis. Section VI will eventually conclude the paper with future extensions.

## II. RELATED WORKS

Evaluating the perception of air pollution in cities today is an important issue. Traditional air quality monitoring methods are expensive to produce, set up, and protect, limiting space security and particle size, and can only be mapped to a limited number of features. A detailed statistical map of the dense air can be obtained from the data gathered by the portable survey structure. This creates a problem because dimensions can be made inside the house (such as a parking silo). Nevertheless, these approaches cannot achieve extreme precision since the device diversity is still too insignificant to capture the entire site and the agenda of moving vehicles changes the daytime directional exposure. In direction to clarify air contaminants from dispersed information, it is essential to evaluate the state of contaminants at any period and in any environment. The existing research on air quality and health tracking is wide-ranging and relevant in many situations. More precisely, the contribution of the works can be outlined posterior to two macro areas. The early neighbourhood is associated with air pollution prediction and early warning systems.

Alimissis et al. [5], assessed binary interpolation methods, an artificial neural network and multiple regression, with statistics from an actual city air quality monitoring system in the Athens metropolitan area of Greece. A statistical study was built on the construction of a timely based air contaminant concentration database to study changes in underground three-dimensional and regular yearly pollutant levels but in real-time assessment and pictures of the situation.

D. Pratiba et al. [6], states that manually saving Google Scholars data can swiftly become a hassle. Therefore, it can be implemented using some code. The Data Journalism Handbook describes several methods for retrieving statistics from the website. It is done by using a web-based Application Program Interface (API), extracting data from a PDF, or browsing a web page. The benefit of discarding is that we can discard it for any other website, even if that website does not have an API to access the rare data. Various tools and techniques can be used for scraping and executing diverse functions. Legibility aids to excerpt the manuscript from a network page. The scraper Wikipedia [7] is a website that allows users to program scraper in some diverse software development idioms. The HyperText Markup Language (HTML) code cast-off to build the website is cast-off by these scrapers, which are Python, Ruby, or any codes [8]. The data must be stored in an organized format before the extracted data can be used for analysis. Gathering Data from the Current Website describes the diverse approaches we can use to store data. Broadcasting records can be saved through the reference or by transferring the folder by themselves. Saving the reference means counting that Uniform Resource Locator (URL). The benefit of exploitation URLs is that the scraper runs ample quicker, uses less bandwidth, and keeps back a lot of disk space [9]. Similarly, the proposed system uses the CPCB websites dataset to web scrap for the future prediction of air quality.

Jiao et al. [10], propose air pollution bases can be classified into fact sources, planar sources, radiation sources, and form sources. Traditional bases of effluence mainly comprise Sulphur dioxide (SO2), Particulate Matter (PM), Nitrogen dioxide (NO2), and Carbon monoxide (CO2) [11]. Over the past years, China has closed some factories and gradually improved the air quality to decrease the rate of air pollution accidents. As of 2019, 4336 aviation surveillance schemes were constructed in cities above county-level cities. So far, the ecological atmosphere tracking station has past data. This data is primarily used for real-time tracking, daily, weekly, and once-a-month reports. Real-time ecological data is also accessible on the environmental monitoring website in China. However, due to the continued growth of air pollution control and study, trends and regulations have attracted the attention of people

with air contaminant awareness predictions [12]. Therefore, in their paper, they have developed a predictive model for the AQI [13] based on Long Short-Term Memory (LSTM). For the air quality data from Shanghai Environmental Monitoring Station from late 2018 to September 2019, they selected 90% as the training set and use the remaining 10% as the test data set. Our proposed paper divides the 73% as training set and 25% as test data.

Y. Lan et al. [14], propose that Long-term air contamination forecast studies are conducted home-based and overseas, and the methods are mostly divided into two groups: traditional distributed models and data-driven models. The classical dispersion model [15] uses contaminant input statistics and weather-related data to set form protection comparations conferring to the calculations of atmospheric dynamic forces and atmospheric chemistry to obtain the altitudinal and time-based dispersal of pollutants. Though, it is very hard to get all of these statistics precisely and fully pretend the atmosphere in the air, so it is difficult to guarantee predictive accuracy. In addition, the computational effort is very high. Therefore, the request for the classical propagation model is inadequate.

W. Zhu et al. [16], states that lately, Recurrent Neural Networks (RNNs) have been used to model air quality and are measured to be beneficial over Artificial Neural Networks (ANN) due to the feature that models connected sequences in time series. [17]. This author proposes modelling covered air quality constraints using Neural Networks (NN), Multiple Regression (MLR), and Recurrent Neural Networks (RNN). The RNN model has excellent performance with active nonlinearity. Long-term and short-term memory models (LSTMs) are special structural types of RNNs. It augments 3 control units. H. Input gate, output gate, oblivion gate. It can solve long-term dependency problems [18] of neural networks. The performance of the LSTM is superior to that of RNNs. The precision of AQI prediction in wide-ranging and unexpected circumstances was high, which was proved in their research.

## III. METHODOLOGY

The overview of various ML and DL algorithms used for the comparison of prediction of air quality is discussed in this section.

### *A. Ridge Regression*

Ridge regression (T. Pu et al., [19]) is added kind of regression process in DNN that is typically measured when there is a high link amongst autonomous variables or model constraints. As the value of the link increases, the least-squares estimation assesses the balanced value. However, if the dataset is very linear, it may be biased. Consequently, generate a bias matrix [22] in the comparison of the ridge regression algorithm. This is a convenient regression technique that works fine with very small datasets, as the models are less likely to overfit.

### *B.Least Absolute Shrinkage and Selection Operation (LASSO)*

The word "LASSO" (L. Zhoul et al., [20]) stands for Least Absolute Shrinkage and Selection Operator. Lasso regression makes predictions according to regulation techniques. It takes precedence done through other regression methods to provide precise predictions. The lasso regression model customs the contraction method. This technique reduces the value of the statistic to the center point [21], as like mean. The lasso regression system proposes a simple and sparse model [23] (that is, a model with few parameters). This is suitable for automating models and data with a high level of multicollinearity, or certain portions of model selection. Variable assortment or constraint removal function uses the L1 regularization method. Since feature selection is performed automatically, it is considered when there are many features in AQI.

### *C.Gated Recurrent Unit (GRU)*

A simpler form of the RNN is called GRU. The typical GRU (Gruber et al., [24]) shape of the cell is shown in figure 1. An emblematic GRU cell has 2 gates. One is the reset gate (r) and the other is the update gate (z). Uses the concealed output value at time t1 and the input time sequence value at time t1. Timestamp t, the hidden output is calculated at time t. This is basically the same as an LSTM cell. The hidden output at time t is calculated using the hidden state at time t1 and the input time series value at the time. This is alike to an LSTM cell. Also, the GRU cell reset gate workflow is suitable for oblivion. Gate of long- and short-term memory cells.

Reset and update gates are used in the hidden layer of the GRU model to modify how state variable calculations are done. In this paradigm, the reset gate is used to disregard the information from earlier states, while the update gate is used to govern the information from earlier hidden states that is carried over into the current state.

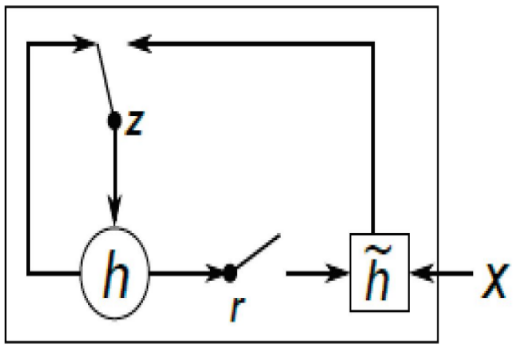


Figure 1 A Simple GRU cell (given by Gruber et al., [24])

*D. Long Short-Term Memory (LSTM)*

LSTMs (Hochreiter et al., [25]) are a type of Time Recursive Network Nerve (RNN) that can process and predict critical events with comparatively long time series intervals and delays. The LSTM differs primarily from the RNN in that it enhances a processor to the process to determine if the data is useful for the air quality prediction. The assembly in which the processor is located is named a cell. There are 4 kinds of gates accessible in classic LSTM. 1st is the input gate, 2nd input modulation gate, 3rd oblivion gate, and 4th Exit gate. An example of an emblematic LSTM cell is demonstrated in Figure 2. The job of the Input Gate is to progress all-new statistics from the open air. The input modulation gate is connected to the recall cell. In the next reiteration, the oblivion gate determines the data to hold and not to hold. Doing so will select the best delay for the input data stream. The calculated one is used as an input to the exit gate. Output gate produces long/short results Term memory cell. Generally, the SoftMax layer is loaded on top of the language's LSTM output layer However, in our model , a LeakyRelu layer is loaded on top of the exit layer of the LSTM cell with sigmoid function. Linear activation function is used in the output layer for networks that predicts air quality is set to Relu. The LeakyReLU activation function has been altered to become leaky. It has the same form as the ReLU, but if some positive values are sufficiently close to zero, it will leak them to 0. If the memory is to be flushed or kept at its current value, the forget gate makes that decision. The output gate, meanwhile, decides whether the memory cell should affect the output at the current time step. X in this procedure = (X1, X2, ..., Xn) is the input time series, H = (H1, H2, ..., Hn) is the hidden state of the memory cell, Y = (Y1, Y2...., Yn) is Output time series. The hidden state (ht) of a memory cell is computed using the following formula (1):

$$ht = H(WhxXt + WhhHt - 1 + bh)\text{--(1)}$$

$$t = WhyYt - 1 + bY\text{----------(2)}$$

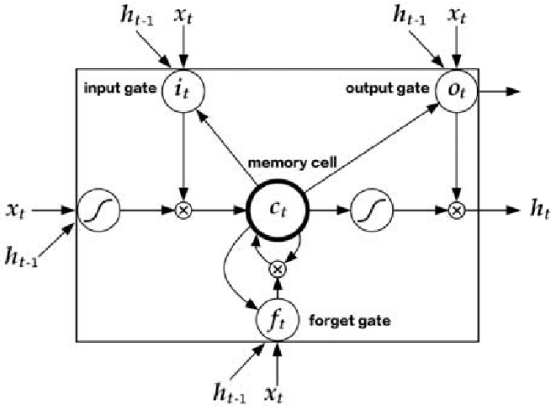


Figure 2 LSTM cell structure (given by Hochreiter et al., [25])

Where t-time gate, hx- hidden time, Xt-procedure time, bh-hidden bias, Yt- time of memory and bY-bias of memory. Here, the weight matrix is shown as "W" and the bias vector is shown as "b". And with the resulting formula (2). We can calculate the hidden state of memory cells through the below equations (3,4,5).

$$ct = ft * ct - 1 + it * g(Wcxxt + Whhht - 1 + Wccct - 1 + bc)\text{-----(3)}$$

$$ot = \sigma(Woxxt + Whhht - 1 + Wocct - 1 + bo)\text{----------(4)}$$

$$ht = ot * h(ct)\text{----------(5)}$$

Where ct-chi square test, * denotes scalar product of two vectors. The standard sigmoid function is represented by σ and given in Eqn. (6), then the sigmoid function is represented by "g". Then the range of the function deviations to [2, 2] and [1, 1].

$$\sigma(x) = 1/1 + e\,x\text{----------(6)}$$

The quadratic loss function is used as the unbiased function conferring to the subsequent equation (7):

$$e = Xn\ i = 1\ (Yt - Pt)\ 2\text{---(7)}$$

Here, the actual output is indicated by "Y" and the predicted AQI value is indicated by "P". Adam Optimizer Applied to perform backpropagation through time to evade local minimums and minimalize training Error. Neural networks were inclined to overfit. Numerous regularization methods have been presented to decrease the overfitting problem. To do this, an efficient way to train a neural network called a dropout has been proposed. Because the dropout method using the iterative property had no effect on RNN (Recurrent Neural Networks (RNN) until 2015.

Among the comparison of four different models, LSTM are selected for the proposed system. Since LSTMs are frequently used for consecutive analysis, they can be

trained to forecast AQI levels for the upcoming hour or even the upcoming month using the historical data gathered by sensors at different weather stations. The information from one of the cities in India's weather stations was used to train the proposed system.

## IV. IMPLEMENTATION AND RESULTS

This section deals with the web scraping technique to scrap the data from websites. ML and DL algorithms are deeply elaborated to find the best-fit algorithm for the proposed model. This section includes an initial examination of the dataset utilized in this research as well as a brief introduction to the investigative methods employed in this work. Fume from automobiles and industry pollutes and makes the air we inhale destructively. Air quality is measured through the scale AQI. This is measured by calculating the average concentration of contaminated particles at standard time intervals. Some Indian metropolises decrease inside the array of the most polluted cities within the international, and the risk of air pollutants is being raised daily. Poor air exceptional in India is now taken into consideration as a tremendous health obligation and the main impediment to economic increase. India's major contaminant productions are from power generation manufacturing, road automobile traffic, soil and road dust, surplus incinerators, power plants, and outdoor incinerators. This work examines air pollution data collected from the CPCB of India. The dataset consists of 12 features with 29,531 rows from 24 diverse Indian cities original dataset. After pre-processing 14402 rows are there in the dataset. Table 1 displays the dataset containing sample observations from 2015 to 2020. Analyzing several chief air pollutants namely PM2.5, PM10, NO2, CO, SO2, and O3, and predicting AQI is the essence of current work with the data variables city and date. The procedural step of the implemented process is shown in figure 3.

### *A. Model Building State*

In this model, data are collected from the open-source Kaggle AQI dataset from the Ministry of Central Government Forest and Climate changes of CPCB website. Using this dataset, we built regression models of DL and ML algorithms. In order to rerun the packages, seeds are set to retain accuracies and can iterate 10 times. First, they read the Comma Separated Value (CSV) file, then the data are represented which has some null values. All the reports are stored in panda's library. Profile reports are created and saved as two different HTML files as original data reports and pre-process data reports using the web scraping technique Web scraping is nothing but extracting the dataset from the website. The entire dataset report has been generated in this report. There are several characteristics such as variables, interactions, correlations, missing values, and the sample. We can access anything from this report. Profile reports are created using panda's profiler.

**Table 1 Sample observations for the Visakhapatnam with air pollutants**

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | O3 | Benzene | Toluene | Xylene | AQI | AQI_Bucket |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **29521** | Visakhapatnam | 2020-06-22 | 33.17 | 108.22 | 5.58 | 42.45 | 27.06 | 13.70 | 0.73 | 13.65 | 34.85 | 3.99 | 10.24 | 2.32 | 95.0 | Satisfactory |
| **29522** | Visakhapatnam | 2020-06-23 | 25.40 | 83.38 | 2.76 | 34.09 | 19.92 | 13.13 | 0.54 | 10.40 | 43.27 | 2.88 | 12.03 | 1.33 | 100.0 | Satisfactory |
| **29523** | Visakhapatnam | 2020-06-24 | 34.36 | 90.90 | 1.22 | 23.38 | 13.12 | 14.45 | 0.56 | 10.92 | 35.12 | 2.99 | 3.15 | 1.60 | 86.0 | Satisfactory |
| **29524** | Visakhapatnam | 2020-06-25 | 13.45 | 58.54 | 2.30 | 21.60 | 13.09 | 12.27 | 0.41 | 8.19 | 29.38 | 1.28 | 5.64 | 0.92 | 77.0 | Satisfactory |
| **29525** | Visakhapatnam | 2020-06-26 | 7.63 | 32.27 | 5.91 | 23.27 | 17.19 | 11.15 | 0.46 | 6.87 | 19.90 | 1.45 | 5.37 | 1.45 | 47.0 | Good |
| **29526** | Visakhapatnam | 2020-06-27 | 15.02 | 50.94 | 7.68 | 25.06 | 19.54 | 12.47 | 0.47 | 8.55 | 23.30 | 2.24 | 12.07 | 0.73 | 41.0 | Good |
| **29527** | Visakhapatnam | 2020-06-28 | 24.38 | 74.09 | 3.42 | 26.06 | 16.53 | 11.99 | 0.52 | 12.72 | 30.14 | 0.74 | 2.21 | 0.38 | 70.0 | Satisfactory |

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | O3 | Benzene | Toluene | Xylene | AQI | AQI_Bucket |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **29528** | Visakhapatnam | 2020-06-29 | 22.91 | 65.73 | 3.45 | 29.53 | 18.33 | 10.71 | 0.48 | 8.42 | 30.96 | 0.01 | 0.01 | 0.00 | 68.0 | Satisfactory |
| **29529** | Visakhapatnam | 2020-06-30 | 16.64 | 49.97 | 4.05 | 29.26 | 18.80 | 10.03 | 0.52 | 9.84 | 28.30 | 0.00 | 0.00 | 0.00 | 54.0 | Satisfactory |
| **29530** | Visakhapatnam | 2020-07-01 | 15.00 | 66.00 | 0.40 | 26.85 | 14.05 | 5.20 | 0.59 | 2.10 | 17.05 | NaN | NaN | NaN | 50.0 | Good |

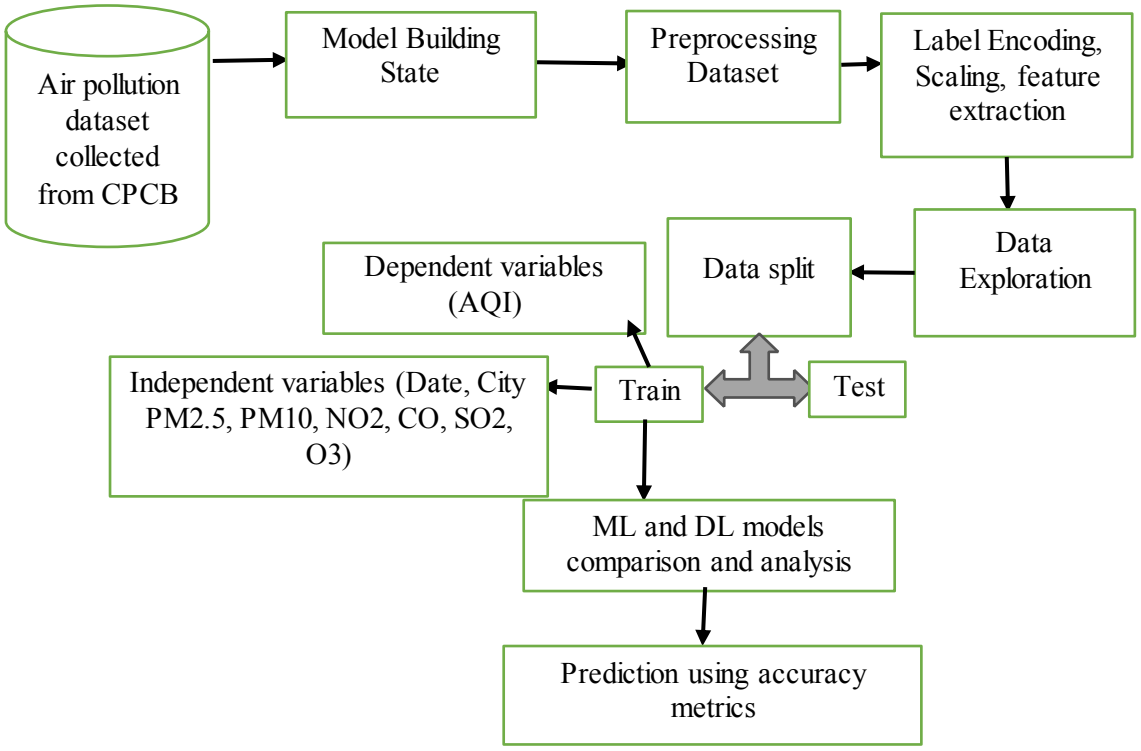


Figure 3 Flowchart of AQPM Model

*B.Pre-processing, Encoding and Scaling*

In this process, we take necessary null data columns and reset them into non-null columns. Figure 4 shows that the dataset consists of non-null values. The Date column will be converted into the year, month and date. Then, unique values are checked in this city column with 26 cities. We are going to utilize label encoding to convert city string object type into numerical. Label encoder from pre-processed library encoded file is saved as a new HTML file for further use. The string object is now converted into int32. Then split the target variable (Y) into the feature variable (X). We are standardizing the whole dataset because each column consists of different kinds of scales. This can avoid feature noises using StandardScaler Library. Bin file is saved for future purposes. Their feature variable is standardized using this library. The X-axis is measured in terms of individual factors and Y-axis is measured in terms of seconds (s).

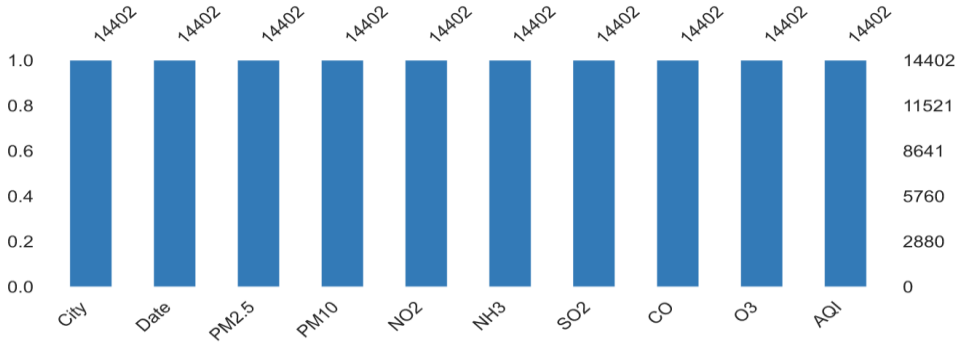


Figure 4 Pre-processed non-null values

*C.MODEL SPLITTING*

The model is split into train and test categories. 75% of data are split into train and 25% of data are split for testing. The training data is divided into independent variables (Date, City, PM2.5, PM10, NO2, CO, SO2, O3) and dependent variables (AQI). The CPCB dataset investigated contains certain parameters. That is AQI and administration agencies custom this constraint to train to warn and predict air quality. Conferring to the National Ambient Air Quality Standards (NAAQS), there are 6 AQI types: Good (0-50), Normal (51-100), Normal (101-200), Bad (201-300), Very Bad (301) --400) and heavy (401-500). Researchers in the area recommend that reducing the input variables reduces the computational rate of modeling and improves predictive performance. In our current work, we used a correlation-based feature assortment procedure to regulate the optimum sum of input variable quantity (contaminants) in the development of the prediction model. Feature selection systems based on statistical correlations calculate the correlation among each pair of input and target variables. The variable quantity that has a robust correlation with the target variables is then filtered for further analysis. Many ML procedures are complex to outliers, so we need to find functionality in the input dataset that see to not track the wide-ranging tendency of that data. The current dataset used correlation-based statistical techniques to detect outliers to categorize outliers. Correlation analysis was performed between AQI properties and other pollutant properties to select key properties. Figure 5 shows the correlation heat map for the AQI.

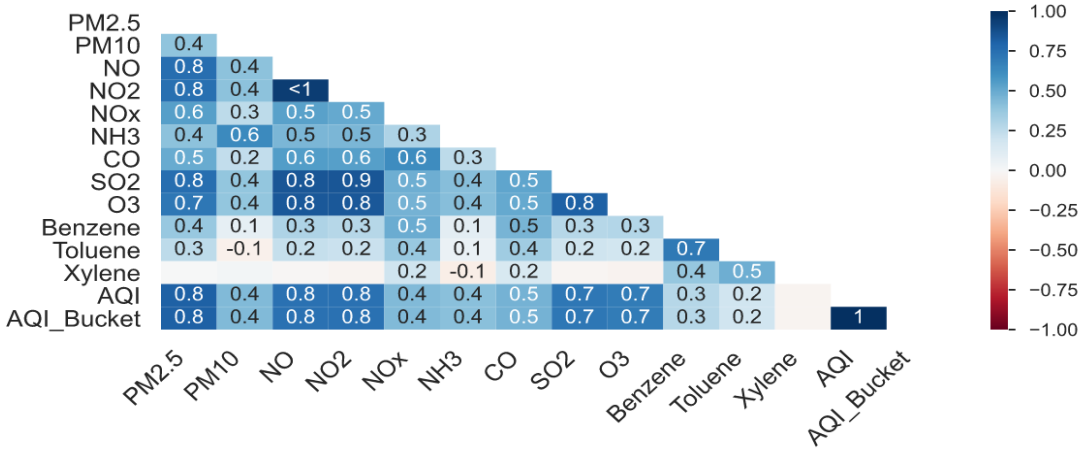


Figure 5 Correlation Heat Map of AQI

*D. BEST FIT REGRESSION MODEL*

The four different regression models are used with advanced ML variations. Ridge, Lasso, Gru, and LSTM are used for comparing three different scores. They are R squared, Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). Two different Recurrent Neural Networks (RNN) algorithms LSTM and GRU are used. We need to reshape the feature column. RNN models are similar to the sequential model. The time factor is given as 1, which means it looks back prior to 1. Finally, dimensions are found in rows, columns, and time factor formats. For example. The output is displayed for 40 hidden LSTM layers activated through GRU. Compiling MSE optimized as Adam epochs for 100, batch size =32. After running the model to 100 epochs, the accuracy score is predicted and acquired three different scores for them. Similarly, all four models run like LSTM. Compared to the other models, LSTM has improved accuracy. The hierarchical data file of

LSTM is exported for future use. Thus, the Kaggle datasets are extracted through Jupiter Notebook data are preprocessed and a regression model are created. CSV files are generated and framed in the panda's data frame.

*E.EXTRACTING PHASE*

These phases are done through three different library requests, beautiful soup, and selenium. First available stations with the available city are acquired. The input parameters are date, City, PM2.5, PM10, NO2, CO, SO2, and O3 and the output will be printed in AQI. Once the stations are acquired in the value list, we pick the already saved file which has label encoding. Label encoding is used to convert the non-numerical data such as city into numerical data. After loading the encoder file, the exact integer value will be assigned. City Delhi will be the first value. Date input will be given and split as date, month, and year. Then, we are converting the integer list into an array. This list will be in the shape where all the data should be in one column. The model is formatted as an array conversion, a standardized file. Tensorflow imports a scalar library through which models are acquired. LSTM is the best fit model which requires parameters such as reshape and time. The model output is predicted and web scraping continues until the prediction is completed. Streamlit packages are used for comparing the four different models. City-wise average AQI acquired is demonstrated in figure 6.

Figure 7 shows the air pollutants (PM 10, PM2.5, CO, NH3, AQI, O3, NO2) levels of the top 5 high average AQI levelled cities. The levels of the Air Quality Index (AQI) are analysed using the independent variable which is given as input. The output dependent variable AQI displays the model encoded values, standardized, predicted values, and acquired website values.

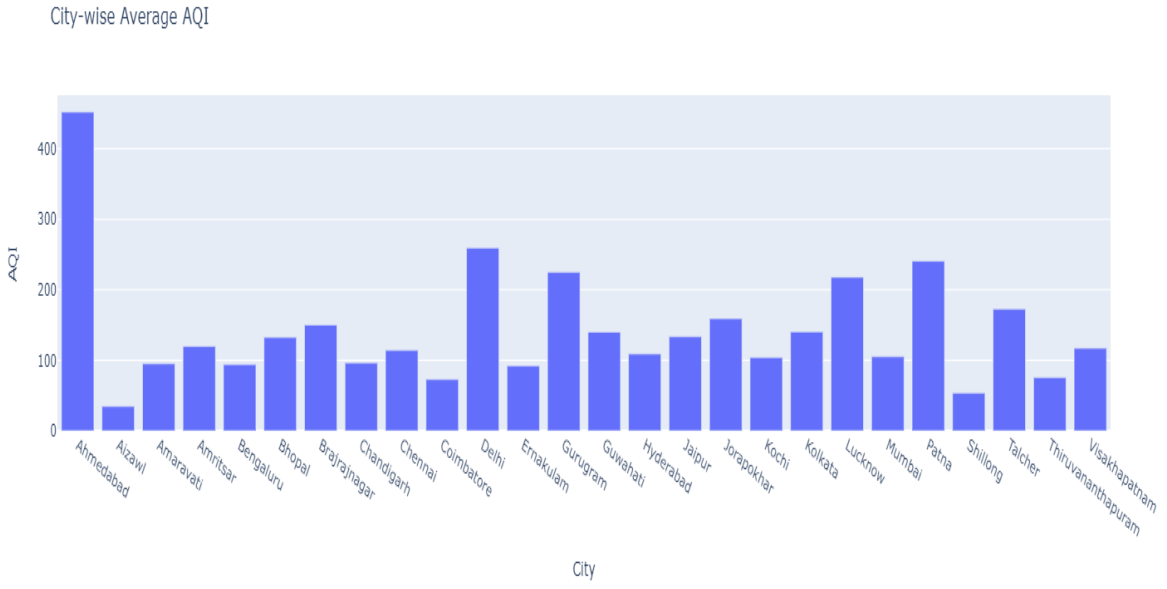


Figure 6 City-wise Average AQI

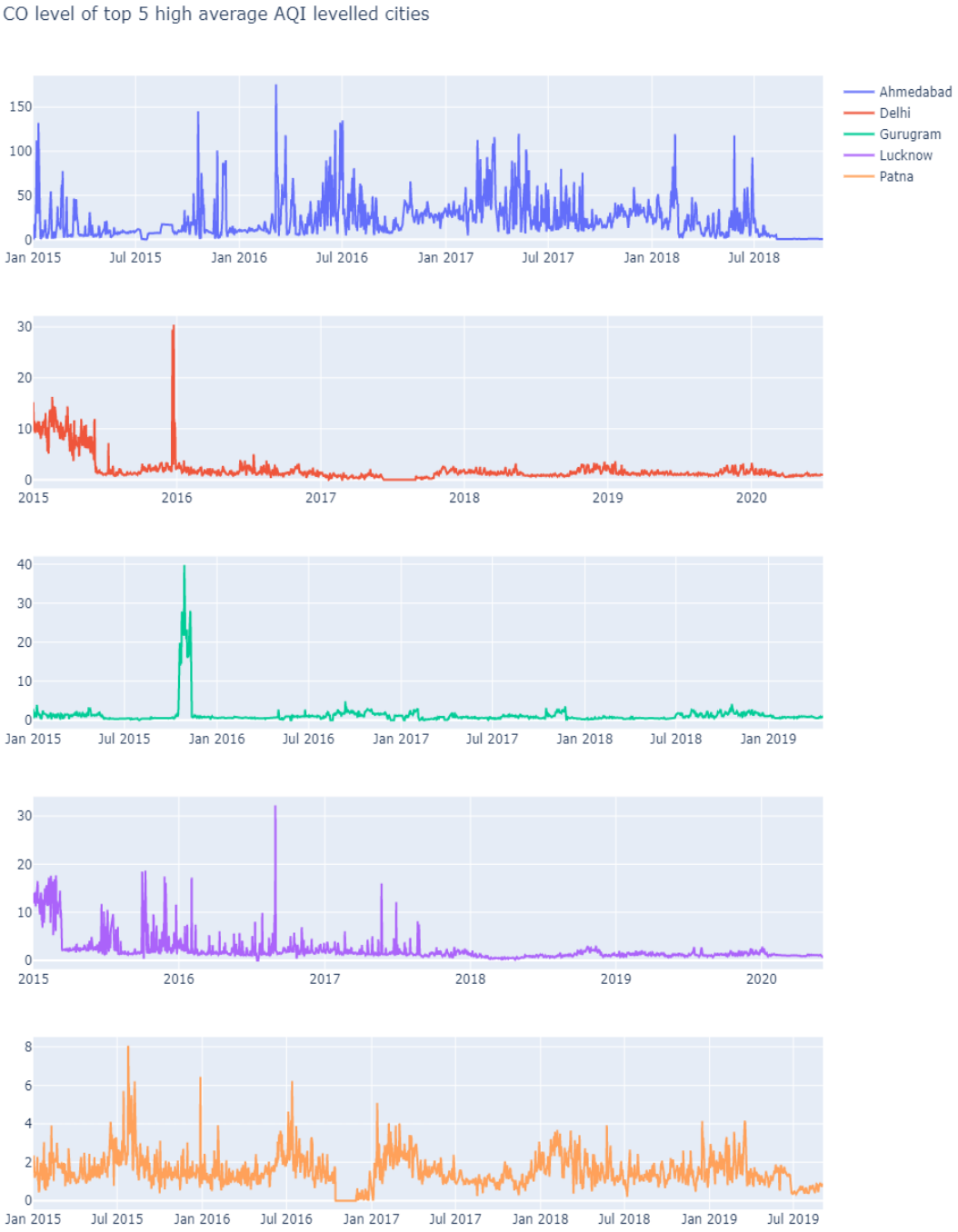


Figure 7 Air pollutants (PM 10, PM2.5, CO, NH3, AQI, O3, NO2) levels of top 5 high average AQI levelled cities

## V. PERFORMANCE METRICS

The accuracy metrics such as R squared, RMSE and MAE are shown in table 2 and represented in figure 8. Mean Absolute Error (MAE) is a measure of the error between pairs of observations. MAE was displayed in terms of seconds. The accuracy of LSTM is high when compared to other models. Examples of Y versus X include predicted and observed times, tracking and initial time comparisons, and measurement and alternative measurement methods. R squared is the square of R and the square root of MSE is RMSE (9) is denoted by the term seconds (s) is expressed in

equation (8). MAE is calculated as in equation (10). Where n denotes sample size.

$$\text{MSE} = \left(\frac{1}{n}\right) * \Sigma\,(\text{actual} - \text{prediction})\,2 - -(8)$$

$$RMSE = \sqrt{MSE} - - - (9)$$

$$MAE = 1/n\sum\frac{predicted - actual}{n}\text{-------}(10)$$

**Table 2 Accuracy Metrics for four different regression model**

| MODELS | R SQUARED | RMSE | MAE |
|---|---|---|---|
| GRU | 0.929938319795973 | 24.17642482824853 | 15.557216517165845 |
| Ridge | 0.892038112254379 | 30.0114541732812 | 20.195209384635305 |
| Lasso | 0.8919448594460742 | 30.02441267356098 | 20.413546694288577 |
| LSTM | 0.9321703453517486 | 23.78820143007658 | 15.438648566309858 |

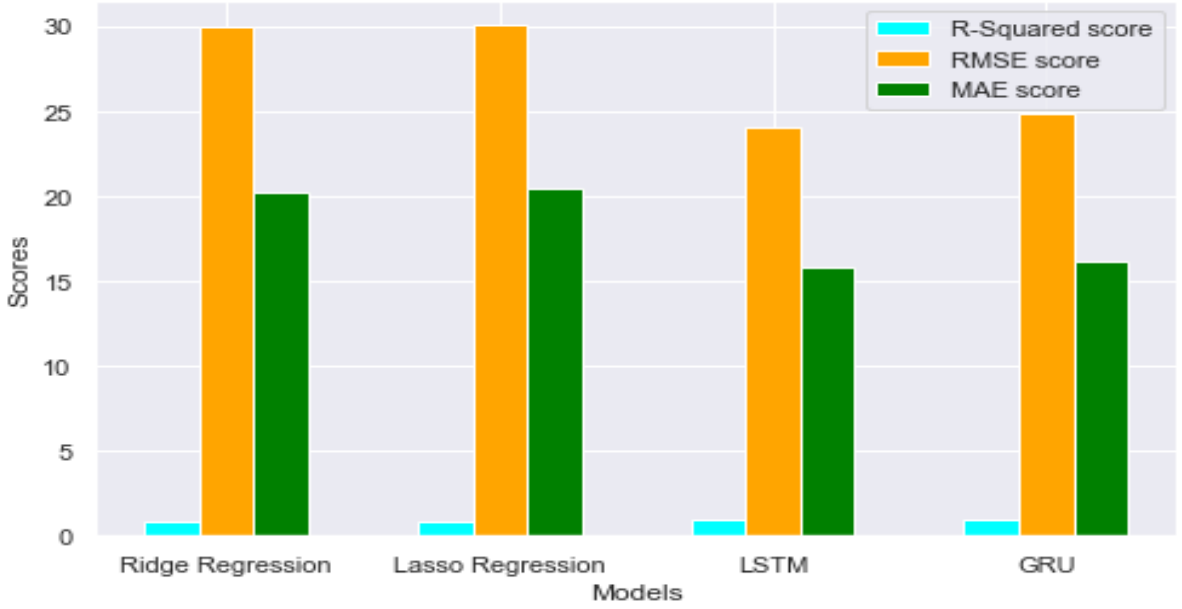


Figure 8 Accuracy Metrics comparison chart

## VI. CONCLUSION AND FUTURE SCOPE

Predicting air quality is a difficult task due to the active atmosphere, volatility, time factor, and spatial and temporal inconsistency of contaminants. The serious significance of air pollution to individuals, creatures, plants, monuments, climate, and the atmosphere requires close tracking and analysis of air quality, specifically in emerging countries. Though, there was not much consideration from investigators on India's AQI predictions. In the current work, air pollution data from 26 Indian cities are being investigated over 6 years. Through the analysis of different ML and DL algorithms, LSTM is the best fit model since it has an R square nearer to 1. We load the dataset, did the pre-processing steps, and found the best fit model through comparison and evaluation of accuracy metrics. The limitations of this work have limited states and cities in the dataset. If the date and time values are not given then the data is not available. The data in the list are appended. If any data is missed then it will display insufficient data or else the model will be encoded, standardized, predict the values and display acquired website values. In the future,

the data with many cities and states can be analyzed. Many other ML and DL models can be compared and analyzed for the most appropriate prediction.

## REFERENCES

[1] Iskandaryan, D., Ramos, F., & Trilles, S. (2020). "Air Quality Prediction in Smart Cities Using Machine Learning Technologies based on Sensor Data: A Review", Applied Sciences, 10(7), 2401. doi:10.3390/app10072401.

[2] Hable-Khandekar, V., & Srinath, P. (2019). "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring". 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). doi:10.1109/iccubea.2017.8463746.

[3] H. Zheng, Y. Cheng and H. Li, "Investigation of model ensemble for fine-grained air quality prediction," in China Communications, vol. 17, no. 7, pp. 207-223, July 2020, doi: 10.23919/J.CC.2020.07.015.

[4] Cynthia Jayapal, Saroja M N, Parveen Sultana, "IoT -Based Real Time Air Pollution Monitoring System", International Journal of Grid and High Performance Computing · October 2019 DOI: 10.4018/IJGHPC.2019100103.

[5] Alimissis A., Philippopoulos K., Tzanis C.G., Deligiorgi D. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmos. Environ.* 2018;**191**:205–213. doi: 10.1016/j.atmosenv.2018.07.058.

[6] D. PRATIBA, A. M.S., A. DUA, G. K. SHANBHAG, N. BHANDARI and U. SINGH, "Web Scraping And Data Acquisition Using Google Scholar," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 277-281, doi: 10.1109/CSITSS.2018.8768777.

[7] Zhou X., Li L., Tong W., Piltner R. *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health.* Elsevier; Amsterdam, The Netherlands: 2020. Sensing air quality: Spatiotemporal interpolation and visualization of real-time air pollution data for the contiguous United States; pp. 169–196.

[8] Wu, Jiahao. (2019). Web Scraping Using Python: A Step By Step Guide. September 2019.

[9] R. Yang, H. Zhou and D. Ding, "Air Quality Prediction Method in Urban Residential Area," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 16-20, doi: 10.1109/ISCID.2018.00010.

[10] Jiao, Yu; Wang, Zhifeng; Zhang, Yang (2019). *[IEEE 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) - Chongqing, China (2019.5.24-2019.5.26)] 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) - Prediction of Air Quality Index Based on LSTM. , (), 17–20.* doi:10.1109/ITAIC.2019.8785602

[11] Y. Lan and Y. Dai, "Urban Air Quality Prediction Based on Space-Time Optimization LSTM Model," 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020, pp. 215-222, doi: 10.1109/ICAIBD49809.2020.9137441.

[12] Lan, Yuxiao; Dai, Yifan (2020). *[IEEE 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD) - Chengdu, China (2020.5.28-2020.5.31)] 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD) - Urban Air Quality Prediction Based on Space-Time Optimization LSTM Model. , (), 215–222.* doi:10.1109/ICAIBD49809.2020.9137441

[13] Abdulrahman Alkandari, Samer Moein, "Implementation of Monitoring System for Air Quality using Raspberry PI: Experimental Study", Indonesian Journal of Electrical Engineering and Computer Science · April 2018 DOI: 10.11591/ijeecs.v10.i1.pp43-49

[14] Y. Lan and Y. Dai, "Urban Air Quality Prediction Based on Space-Time Optimization LSTM Model," 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020, pp. 215-222, doi: 10.1109/ICAIBD49809.2020.9137441.

[15] A. Barve, V. Mohan Singh, S. Shrirao and M. Bedekar, "Air Quality Index forecasting using parallel Dense Neural Network and LSTM cell," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154069.

[16] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Cooccurrence feature learning for skeleton-based action recognition using regularized deep LSTM networks, Proceedings of the AAAI, 2016, 3697–3703

[17] Qi, Z., Wang, T., Song, G., Hu, W., Li, X., Zhang, Z.M.: Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality",2018, IEEE Transaction of Knowledge and Data Engineering. 30, 2285–2297.

[18] K. Nagrecha et al., "Sensor-Based Air Pollution Prediction Using Deep CNN-LSTM," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020, pp. 694-696, doi: 10.1109/CSCI51800.2020.00127.

[19] T. Pu, H. Cai, G. He, Y. Luo and M. Wu, "An air quality prediction model based on deep learning and wavelet analysis considering the COVID-19 pandemic factors," 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC), 2021, pp. 1-2, doi: 10.1109/IPCCC51483.2021.9679440.

[20] L. Zhoul, M. Chenl and Q. Ni, "A hybrid Prophet-LSTM Model for Prediction of Air Quality Index," 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 595-601, doi: 10.1109/SSCI47803.2020.9308543.

[21] Kok, I., Simsek, M.U., & Özdemir, S. (2017). A deep learning model for air quality prediction in smart cities. 2017 IEEE International Conference on Big Data (Big Data), 1983-1990.

[22] Joby, P. P. "Expedient information retrieval system for web pages using the natural language modeling." Journal of Artificial Intelligence 2, no. 02 (2020): 100-110.

[23] Chen, Joy Iong Zong, and Lu-Tsou Yeh. "Graphene based Web Framework for Energy Efficient IoT Applications." Journal of Information Technology 3, no. 01 (2021): 18-28.

[24] Gruber, Nicole and Jockisch, Alfred, ''Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text'', Frontiers in Artificial Intelligence, doi: 3. 10.3389/frai.2020.00040.

[25] Hochreiter, Sepp and Schmidhuber, Jürgen."Long Short-term Memory'',Neural computation, Vol. 9.2019, pp. 1735-80 , doi:10.1162/neco.1997.9.8.1735.